INSTITUTE FOR DEFENSE ANALYSES

Metrics for Assessing Underwater Demonstrations for Detection and Classification of Unexploded Ordnance

(Presentation)

Jacob B. Bartel
Shelley M. Cazares

March 2021

IDA Document NS D-21603

Log: H 20-000091

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication

This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Project AM-2-1528, "Assessment of Traditional and Emerging Approaches to the Detection and Classification of Surface and Buried Unexploded Ordnance (UXO)," for the Director, Environmental Security Technology Certification Program (ESTCP) and Strategic Environmental Research and Development Program (SERDP), under the Deputy Assistant Secretary of Defense (Environment). The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

For More Information

Shelley M. Cazares, Project Leader
scazares@ida.org, 703-845-6792

Leonard J. Buckley, Director, Science and Technology Division
lbuckley@ida.org, 703-578-2800

Copyright Notice

Rigorous Analysis | Trusted Expertise | Service to the Nation

Executive Summary

Receiver operating characteristic (ROC) curves are often used to assess the performance of detection and classification systems. The Strategic Environmental Research and Development Program/Environmental Security Technology Certification Program (SERDP/ESTCP) is sponsoring the development of novel systems for detecting and classifying unexploded ordnance (UXO) in underwater environments. SERDP is also sponsoring underwater testbeds to demonstrate the performance of these novel systems in relevant environments. The Institute for Defense Analyses is currently designing and implementing the scoring process for these underwater demonstrations to address the subtleties of ROC curve interpretation. This presentation provides an overview of the main considerations for ROC curve parameter selection when scoring underwater demonstrations for detecting and classifying UXO.

Metrics for Assessing Underwater Demonstrations for Detection and Classification of Unexploded Ordnance

Jacob Bartel

Shelley Cazares

April 2021

Institute for Defense Analyses

4850 Mark Center Drive ● Alexandria, Virginia 22311-1882

jbartel@ida.org; 703 845 2172

This briefing discusses metrics for assessing underwater demonstrations for the detection and classification of Unexploded Ordnance (UXO). This is work that has been done at the Institute for Defense Analyses (IDA). IDA is a federally funded research and development center that provides scientific and technical analyses for the U.S. government on national security issues.

Unexploded Ordnance (UXO) versus Clutter

- UXO are duds: munitions that were previously armed and fired but did not explode
- UXO can still pose a risk of detonation, even decades later
- Millions of acres of land in the continental U.S. are contaminated with UXO, due to their previous uses as military training camps and test ranges

UXO (TOI)

VS.

Clutter (Non-TOI)

IDA | 1

Unexploded Ordnance or UXO refers to munitions that were previously armed and fired, but did not explode. These UXO duds can still pose a risk of detonation, even decades later. Although most people think of UXO in the context of World War II era bombs in Europe, or Vietnam era bombs in Southeast Asia, millions of acres of land in the *continental United States* are contaminated with UXO, due to their previous uses as military training camps and test ranges.

Introduction

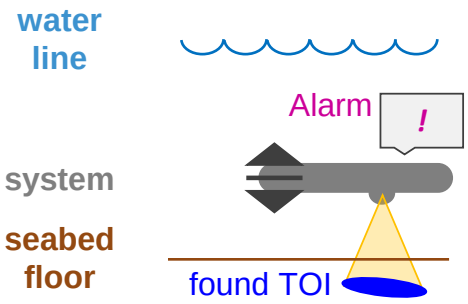
- SERDP sponsors testbeds to demonstrate novel systems for the detection and classification of unexploded ordnance (UXO) in underwater environments
- The Institute for Defense Analyses (IDA) previously scored SERDP's and ESTCP's terrestrial demonstrations, and is now involved in underwater scoring
- Designing and scoring these 'blind tests' is a complicated process
- Interpreting the scores properly reveals subtleties in constructing **Receiver-Operating Characteristic (ROC) curves**

SERDP = Strategic Environmental Research and Development Program
ESTCP = Environmental Security Technology Certification Program

The Strategic Environmental Research and Development Program (SERDP) sponsors testbeds to demonstrate novel systems for the detection and classification of UXO in underwater environments. IDA, having previously led the scoring of *land based* UXO detection and classification system tests, is now involved in *underwater* test scoring. Designing and scoring these blind tests is a complicated process. Interpreting the scores properly reveals some subtleties even in familiar places, such as when constructing Receiver-Operating Characteristic (ROC) curves.

Problem Description

- UXO remediation efforts require information on location and state of targets at sites
- Novel technologies for UXO detection + classification in underwater sites are being developed:
  - Acoustic, EMI, etc. based sensors for detection
  - Human or machine learning based classification
    - Differentiate between a Target of Interest (TOI) such as a UXO vs. a non-TOI such as clutter
- First tests in a **relevant environment** have begun
- First **evaluations** of those tests have also started



EMI = Electromagnetic Induction

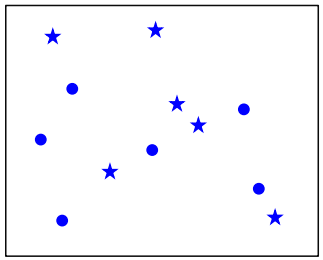
The problem the UXO remediation community is tackling is how to best remediate or manage a site suspected of containing UXO underwater. While different sites may approach the remediation/management problem differently, one requirement that will always be necessary is *information*. SERDP's goal is to promote the creation of systems that are capable of detecting and classifying UXO in an underwater environment, and to evaluate those systems. These novel technologies often revolve around acoustic or electromagnetic induction (EMI) sensors for detection, with a human or machine learning based classification procedure. The primary goal of a system is to differentiate between targets of interest (TOIs), which are items that could be dangerous like UXO, versus non-TOIs such as clutter. An example of a TOI would be an unexploded 155mm howitzer round, and an example of a non-TOI would be a discarded scuba tank.

These technologies are not yet mature, as underwater testing in a relevant environment has just begun. Simultaneously, the first evaluations of those tests are being performed by IDA. Currently, these technologies are at the 'blind test' stage, performed at controlled testing sites.

Blind Test Setup

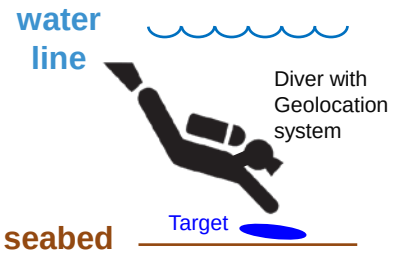
1. Testbed design

(Bird's Eye View of Test Site)



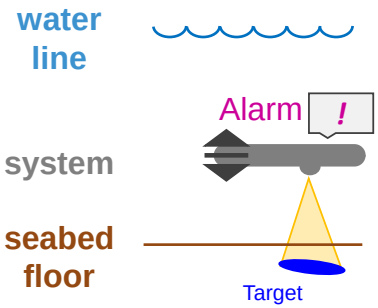
2. Target emplacement

(Side View of Test Site)



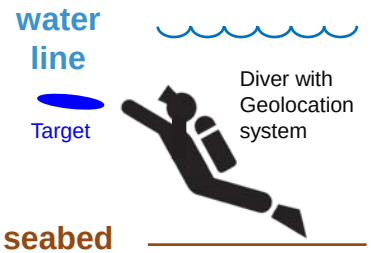
3. Blind test of system

(Side View of Test Site)



4. Ground Truth collection and removal

(Side View of Test Site)



1. Testbed is designed to randomly distribute targets (TOI and non-TOI)
2. Divers emplace targets (buried or proud) at testbed
3. Demonstrators (blind to ground truth) deploy their system within the test area and collect data
4. Divers collect ground truth (geolocation, orientation, burial depth, etc.) and remove targets
5. Demonstrators create a 'call list' detailing what they detected and classified

Here is a general overview of the type of blind test that is used to evaluate one of these novel systems. First, a testbed is designed by selecting TOI and non-TOI to emplace in a roughly random pattern. TOI typically consist of inert munitions or surrogate munitions, and non-TOIs may be any number of different clutter objects that are typical in an underwater environment, like discarded scuba tanks or crab pots. Next, these targets are emplaced in the test area by divers, who may either bury the target in the sediment of the test area, or leave the target proud on the surface, depending on the testbed design specifications. Once all targets are emplaced, a demonstrator will deploy their system in the test area, scanning to collect data that can later be analyzed, in order to detect and classify the targets, offline. Finally, divers once again re-enter the test area to collect ground truth (consisting of coordinates of the emplaced targets, their orientations, and burial depths) and then remove the emplaced items. Afterwards, offline, the demonstrators of the UXO system create a ‘call list’ which is a list of detected items that they identified as possible targets, ranked by the likelihood of each item being a TOI. The ground truth is kept sequestered, and so the demonstrators have to create this list ‘blind’, using only the data they collected during the test. This ‘call list’ is then sent to the scoring team to be scored, such as IDA.

Two Underwater Demonstrations Sites to Date

Sequim Bay, WA

https://mapswire.com/countries/united-states/

Boston Harbor, MA

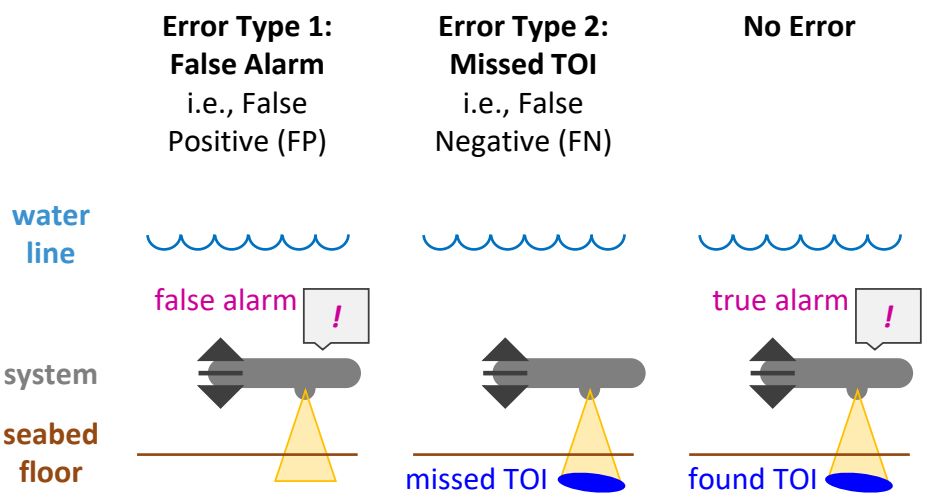
To date, there are two sites that have supported demonstrations—one in Boston Harbor, MA, and one in Sequim Bay, WA.

What kinds of scores are calculated?

After ground truth is collected and the demonstrator submits their **call list** to the scoring team…

Two types of scores are needed:

1. **Probability of False Alarm (Pfa):**
   describes how often the system creates a False Positive (FP), i.e., a False Alarm
2. **Probability of Detection (Pd,c):**
   describes how often the system avoids a False Negative (FN), i.e., a missed Target of Interest (TOI)



False Positives and Negatives trade off of each other:
As one count gets better, the other can get worse

IDA

After the blind test is over, and ground truth has been collected by the diving team, and the demonstrator has turned over their call list to the scoring team, it is possible to begin scoring the demonstration. Broadly speaking, in order to score a demonstration, two types of scores are needed—a score related to the Probability of False Alarm (Pfa), or how often a system creates a False Positive (FP), and a score related to the Probability of Detection and Correct Classification (Pd,c), or how often the system finds the TOI—that is, how often it avoids a False Negative (FN) or misses the TOI. False Positives and False Negatives generally trade off of each other: as one gets better, the other may get worse.

On the right-hand side there is an illustration that shows the two types of error types described here. If the system believes it has detected a TOI, but there turns out *not* to be a TOI at that location, this is a False Positive (or False Alarm). If the system scans over a TOI but does not identify it as a TOI, then that is a False Negative. A True Positive (or True Alarm) occurs when the system correctly identifies the location of a TOI.

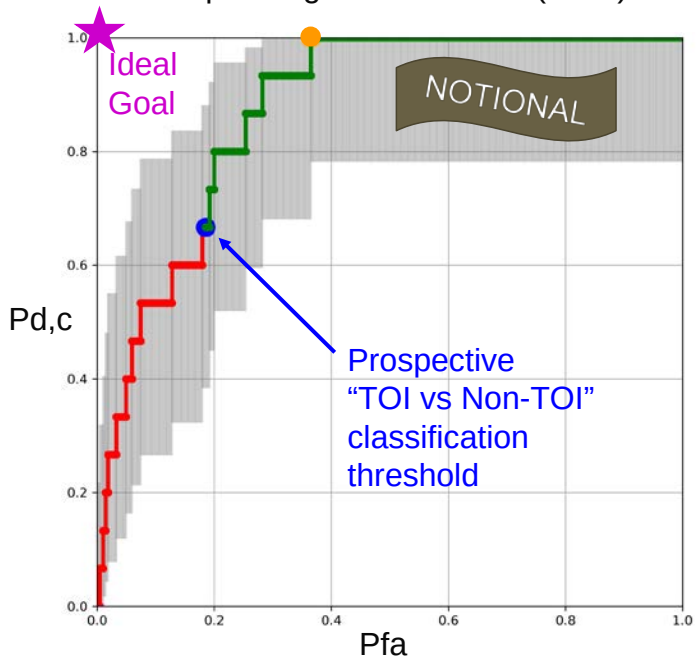
ROC Curve

Once you have Pd,c and Pfa from a blind test you can create a ROC curve…

$$\text{Pd, c} = \frac{\text{\# True Alarms}}{\text{\# True TOIs}}$$

$$\text{Pfa} = \frac{\text{\# False Alarms}}{\text{\# True Non} - \text{TOIs}}$$

Receiver-Operating Characteristic (ROC) Curve



Finally, once you have calculated the Pd,c and Pfa of each item on the demonstrator's call list, based on the ground truth, you can simply create a ROC Curve, seen on the right-hand side. This is a *notional* ROC curve from a notional system, based on notional data.

ROC curves help us visualize the tradeoff between False Negatives and False Positives– between TOI Misses and False Alarms:

- Pd,c is plotted on the Y axis. Pd,c indicates how well the system can find a TOI, or avoid a False Negative. Pd,c ranges from 0 to 1, with higher values considered better.
- Pfa is plotted on the X axis. Pfa is an indication of how often the system generates a False Positive. Pfa also ranges from 0 to 1, but this time, lower values are considered better.

Each point on the ROC curve corresponds to a different classification threshold. A vertical grey line is drawn through each point, to indicate the 95% confidence interval around that point's Pd,c value on the Y axis. Each point's confidence interval was calculated with the beta distribution approximation to the binomial distribution, with no adjustments for multiple comparisons.

The pink star indicates the ideal goal of a detection and classification system—a perfect Pd,c of 1 and a perfect Pfa of 0. In such a case, 100% of the TOIs are found (no False Negatives), with zero False Alarms (no False Positives). That is, ideally, a ROC curve would touch the upper-left corner of ROC space. In reality, though, most ROC curves never reach this ideal goal, and look like this, instead. Still, developers often seek to create systems that get close to the upper-left corner of ROC space.

The blue dot is the *prospective* classification threshold—the threshold selected by the demonstrators *during* the blind test to differentiate between TOI and Non-TOI targets, without any knowledge of ground truth.

ROC Curve

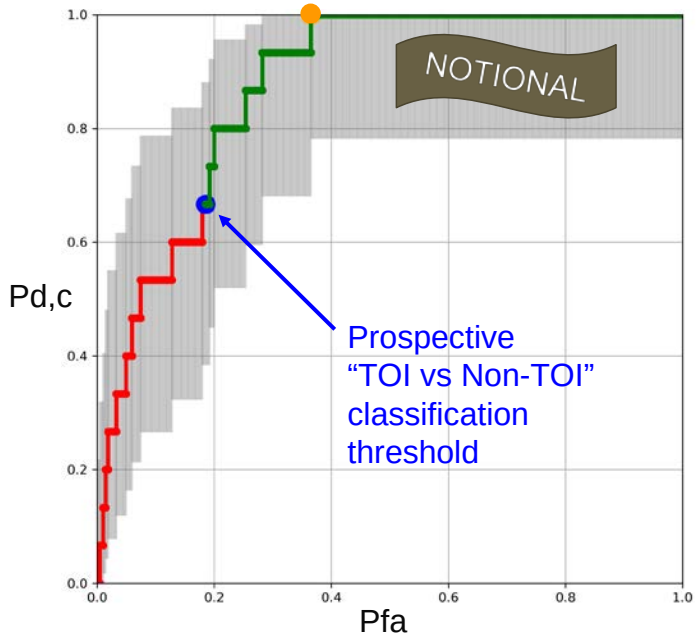
Once you have Pd,c and Pfa from a blind test you can create a ROC curve…

$$\text{Pd, c} = \frac{\text{\# True Alarms}}{\text{\# True TOIs}}$$

$$\text{Pfa} = \frac{\text{\# False Alarms}}{\text{\# True Non} - \text{TOIs}}$$

…But wait!

Receiver-Operating Characteristic (ROC) Curve



But that’s not the end of the story! It turns out, the definitions of Pd,c and Pfa are actually much more complicated than this.

Equations

Y-Axis Metrics

Textbook definition

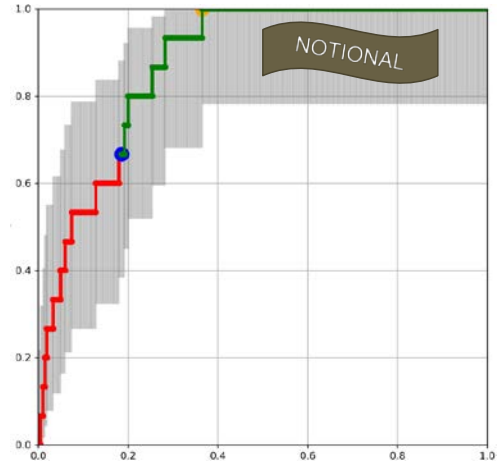
IDA definition

- $\text{Pd, c} = \frac{\text{\# True Alarms}}{\text{\# True TOIs}} \approx \frac{\text{\# True Alarms}}{\text{\# Emplaced TOIs}}$

There are <u>many</u> valid definitions for the ROC curve x-axis metric, all based on underlying assumptions

X-Axis Metrics

- $\text{Pfa1} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Emplaced Non-TOIs}}$
- $\text{Pfa2} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections}}$
- $\text{Pfa3} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections} - \text{\# Emplaced TOIs}}$
- $\text{Pfa4} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections} - \text{\# Found TOIs}}$
- $\text{FP} = \text{\# False Alarms}$
- $\text{FAR} = \frac{\text{\# False Alarms}}{\text{Test Area}}$



IDA 9

Over the course of UXO system evaluations that IDA has been involved in, we have documented a number of subtleties involving ROC curves for UXO detection and classification. There are actually many valid definitions for the x and y ROC curve axes that could be used, depending on what story you'd like to tell. In each of these equations, we present the 'textbook' definition of the quantity we're referring to, and then the IDA definition, which is designed to be workable given the real-world data that we can actually acquire in these tests.

Starting from the top: the definition of Pd,c is fairly standard, with one modification to the textbook definition in the IDA definition: we assume that the number of true TOIs is approximately equal to the number of emplaced TOIs we have the divers put down in the test area. In other words, we are assuming that there are no native UXO or UXO-like objects in the test area. This assumption must be made, because otherwise you cannot define the actual number of True TOIs.

Equations

Y-Axis Metrics

Textbook definition

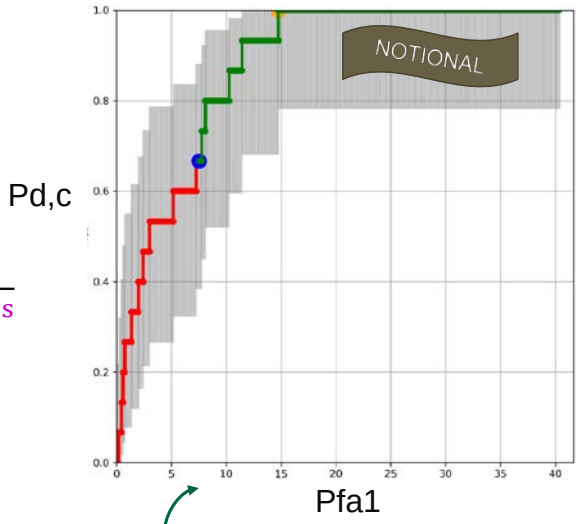
IDA definition

- $\text{Pd, c} = \frac{\text{\# True Alarms}}{\text{\# True TOIs}} \approx \frac{\text{\# True Alarms}}{\text{\# Emplaced TOIs}}$

X-Axis Metrics

- $\text{Pfa1} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Emplaced Non-TOIs}}$
- $\text{Pfa2} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections}}$
- $\text{Pfa3} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections} - \text{\# Emplaced TOIs}}$
- $\text{Pfa4} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections} - \text{\# Found TOIs}}$
- $\text{FP} = \text{\# False Alarms}$
- $\text{FAR} = \frac{\text{\# False Alarms}}{\text{Test Area}}$

There are many valid definitions for the ROC curve x-axis metric



Note the units and scale!

Next, on the X-Axis, we see four different possible definitions for Pfa. Pfa, in the textbook definition, is a metric where the number of False Alarms a system creates (numerator) is normalized to the number of True Non-TOIs (or clutter objects) that are present at the site (denominator). However, the ‘number of True non-TOIs’ is a number that is difficult to capture, and so approximations or replacements must be made for the denominator.

First, Pfa #1, assumes that the only non-TOIs that are in the test area are the ones that the divers emplaced as part of the test. This sounds good—however the problem is that often there are many Non-TOIs that occur *naturally* in test bed areas—such as trash, rocks, anchors, and so on. So, in reality, the number of emplaced Non-TOIs in a test area may actually be quite far off from the *true* number of non-TOIs.

Equations

Y-Axis Metrics

Textbook definition

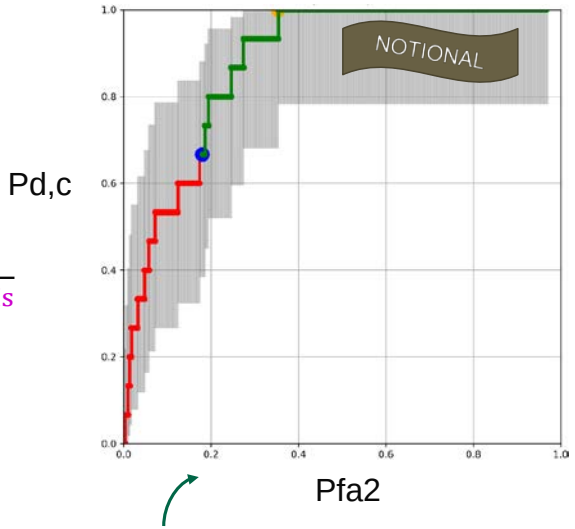
IDA definition

- $\text{Pd, c} = \frac{\text{\# True Alarms}}{\text{\# True TOIs}} \approx \frac{\text{\# True Alarms}}{\text{\# Emplaced TOIs}}$

There are <u>many</u> valid definitions for the ROC curve x-axis metric

X-Axis Metrics

- $\text{Pfa1} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Emplaced Non-TOIs}}$
- → $\text{Pfa2} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections}}$
- $\text{Pfa3} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections} - \text{\# Emplaced TOIs}}$
- $\text{Pfa4} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections} - \text{\# Found TOIs}}$
- $\text{FP} = \text{\# False Alarms}$
- $\text{FAR} = \frac{\text{\# False Alarms}}{\text{Test Area}}$



IDA 11

This takes us to Pfa #2, where the number of False Alarms (numerator) is normalized to the number of *detections* the UXO system found when surveying the test area (denominator). In effect, this Pfa metric shows how much the classification part of the UXO system cut down on its own detections. In other words, at each point in the ROC curve, this Pfa metric tells you how many false positives are required to achieve a certain Pd,c value, as a fraction of the total number of detections the system made.

Equations

Y-Axis Metrics

Textbook definition

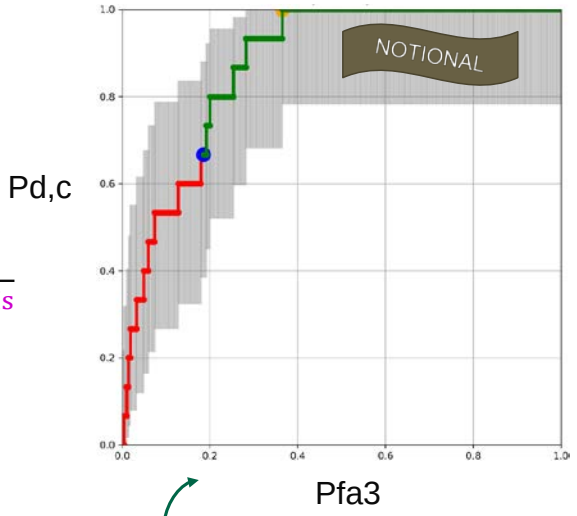
IDA definition

- $\text{Pd, c} = \frac{\text{\# True Alarms}}{\text{\# True TOIs}} \approx \frac{\text{\# True Alarms}}{\text{\# Emplaced TOIs}}$

X-Axis Metrics

- $\text{Pfa1} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Emplaced Non-TOIs}}$
- $\text{Pfa2} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections}}$
- → $\text{Pfa3} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections} - \text{\# Emplaced TOIs}}$
- $\text{Pfa4} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections} - \text{\# Found TOIs}}$
- $\text{FP} = \text{\# False Alarms}$
- $\text{FAR} = \frac{\text{\# False Alarms}}{\text{Test Area}}$

There are many valid definitions for the ROC curve x-axis metric



Note the units and scale!

However, some detections are not False Positives—and so the X axis of the ROC curve shouldn't "ding" the system for detecting a UXO or other TOI-like object. To account for this, Pfa #3 removes the number of emplaced TOIs from the denominator, so that the denominator is now the maximum number of False Positives that the system achieved in the demonstration. This Pfa metric is one of the more intuitive and widely used definitions for Pfa.

Equations

Y-Axis Metrics

Textbook definition

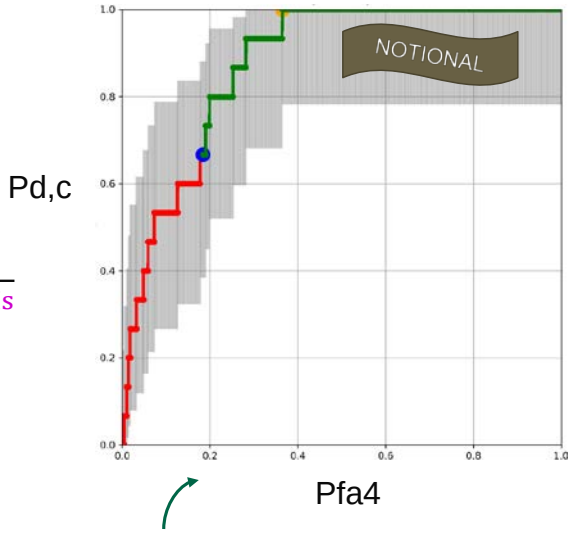
IDA definition

- $\text{Pd, c} = \frac{\text{\# True Alarms}}{\text{\# True TOIs}} \approx \frac{\text{\# True Alarms}}{\text{\# Emplaced TOIs}}$

X-Axis Metrics

- $\text{Pfa1} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Emplaced Non-TOIs}}$
- $\text{Pfa2} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections}}$
- $\text{Pfa3} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections} - \text{\# Emplaced TOIs}}$
- → $\text{Pfa4} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections} - \text{\# Found TOIs}}$
- $\text{FP} = \text{\# False Alarms}$
- $\text{FAR} = \frac{\text{\# False Alarms}}{\text{Test Area}}$

There are many valid definitions for the ROC curve x-axis metric



Note the units and scale!

Finally, the number of *potential* False Positives at every point in the ROC curve does not stay constant—because as one goes along the ROC curve from bottom-left to top-right, some TOIs are detected and classified correctly. So, to account for this, Pfa #4 only subtracts, from the number of detections in the denominator, the number of ***found*** TOIs at each point in the ROC curve. This makes the denominator a variable, not a constant, as others were.

Equations

Y-Axis Metrics

Textbook definition

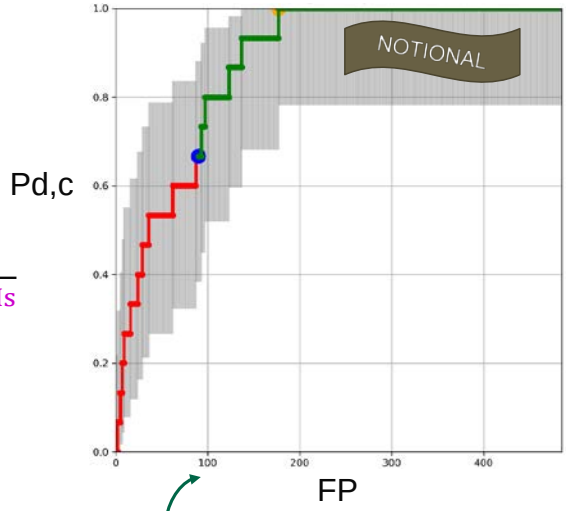
IDA definition

- $\text{Pd, c} = \frac{\text{\# True Alarms}}{\text{\# True TOIs}} \approx \frac{\text{\# True Alarms}}{\text{\# Emplaced TOIs}}$

There are <u>many</u> valid definitions for the ROC curve x-axis metric

X-Axis Metrics

- $\text{Pfa1} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Emplaced Non-TOIs}}$
- $\text{Pfa2} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections}}$
- $\text{Pfa3} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections} - \text{\# Emplaced TOIs}}$
- $\text{Pfa4} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections} - \text{\# Found TOIs}}$
- $\text{FP} = \text{\# False Alarms}$
- $\text{FAR} = \frac{\text{\# False Alarms}}{\text{Test Area}}$



Note the units and scale!

Pfa-type metrics are very commonly used in ROC curves. However, there are additional alternatives, including simply using the number of False Alarms.

Equations

Textbook definition

IDA definition

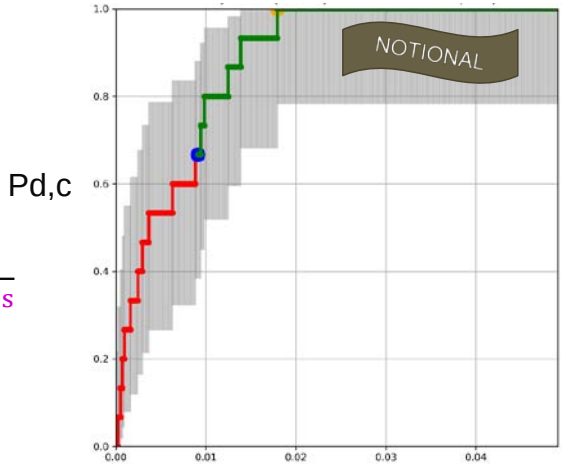
Y-Axis Metrics

- $\text{Pd, c} = \frac{\text{\# True Alarms}}{\text{\# True TOIs}} \approx \frac{\text{\# True Alarms}}{\text{\# Emplaced TOIs}}$

There are <u>many</u> valid definitions for the ROC curve x-axis metric

X-Axis Metrics

- $\text{Pfa1} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Emplaced Non-TOIs}}$
- $\text{Pfa2} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections}}$
- $\text{Pfa3} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections} - \text{\# Emplaced TOIs}}$
- $\text{Pfa4} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections} - \text{\# Found TOIs}}$
- $\text{FP} = \text{\# False Alarms}$
- $\text{FAR} = \frac{\text{\# False Alarms}}{\text{Test Area}}$



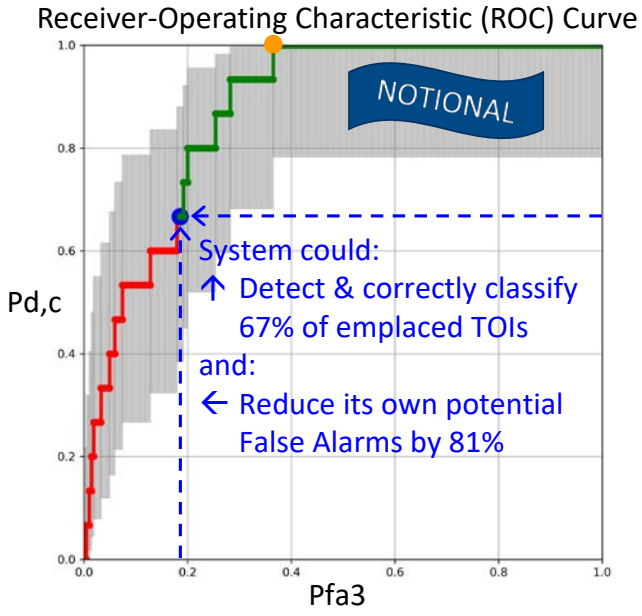
Note the units and scale!

IDA 15

Or, alternatively, the number of False Alarms normalized by the physical size of the test area. We will touch on these in the next few slides.

Example: Notional ROC Curve A

- Y Axis: $\text{Pd, c} = \frac{\text{\# True Alarms}}{\text{\# True TOIs}} \approx \frac{\text{\# True Alarms}}{\text{\# Emplaced TOIs}}$
- X Axis: $\text{Pfa} = \frac{\text{\# False Alarms}}{\text{\# True Non-TOIs}} \approx \frac{\text{\# False Alarms}}{\text{\# Detections} - \text{\# Emplaced TOIs}}$
- **Pfa3** summarizes how well the system's Classification step corrected the potential False Alarms from its Detection step:
  - Pro: Easy to compare between demonstrations, since Pfa ranges from 0 to 1
  - Con: Difficult to compare between systems, since Pfa (as defined here) is a relative measure comparing different steps of the same system

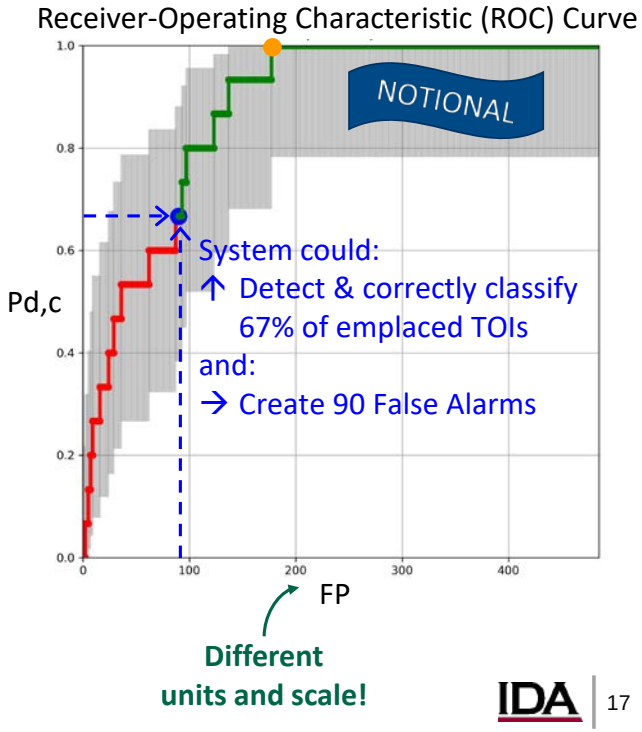


Now we will look at a few examples of ROC curves using these different x-axes metrics to show how the story changes with each one. First, we'll look at a *Pfa-lik*e metric (Pfa #3 from the last slide). This Pfa metric summarizes how well the system's classification step corrected the potential false alarms from its own detection step. In the ROC curve, the blue dot points out the demonstrator's prospective "TOI vs. Non-TOI" classification threshold. You can see that at the blue dot, the system was able to correctly detect and classify 67 percent of the emplaced TOIs (vertical axis), while reducing its own potential False Alarms by 81 percent (horizontal axis).

The benefits of this metric are that all x-axis values range from 0 to 1, allowing for easy comparisons between different demonstrations. However, this metric also makes it difficult to compare between different systems, since Pfa is a relative measure comparing different steps of the same system (detection vs. classification). Some systems prefer to have a lot of detections, that they then whittle down with their classification step. Other systems prefer to have a less sensitive detection step, so that their classification step has less work to do. In this Pfa metric, the first system's ROC curve *shape* would look better than the second system's, though their *final* results—the locations of their blue dots—may be identical.

Example: Notional ROC Curve B

- Y Axis: $\text{Pd, c} = \frac{\text{\# True Alarms}}{\text{\# True TOIs}} \approx \frac{\text{\# True Alarms}}{\text{\# Emplaced TOIs}}$
- X Axis: $\text{FP} = \text{\# False Alarms}$
- **FP** is a final count of the system's False Alarms after both Detection and Classification:
  - Pro: Easy to compare between systems, since FP is an absolute count of False Alarms
  - Pro: FP count can be converted to real-world remediation cost ($/dig)
  - Con: Not easy to compare between demonstrations, since test area sizes vary

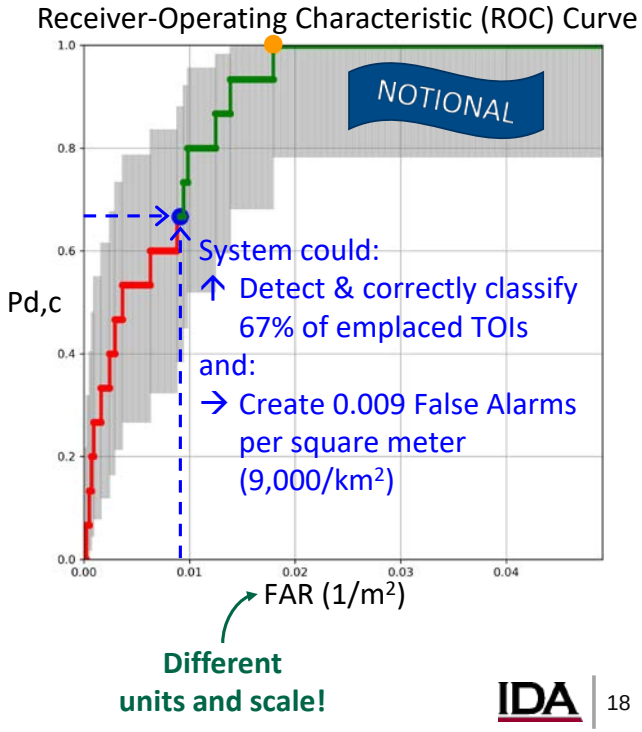


One way to get around the comparability issue that the Pfa metric on the last slide has, is to simply plot the number of *False Positives* on the x-axis (the numerator, all by itself). Plotting this number results in an easy comparison between systems, since systems that prefer either a stringent detection or classification step wouldn't be artificially 'dinged' for choosing one approach over the other. Additionally, in a real-world remediation project, the number of False Positives is directly related to the remediation cost—i.e. each 'dig' when removing a suspected UXO costs money.

However, the False Positive metric does not make it easy to compare between different demonstrations, since the area of testbeds can differ significantly for different sites. For instance, a ROC curve from a demonstration in a large test area may look worse than a demonstration in a small test area, because the larger test area may provide more opportunity for False Alarms.

Example: Notional ROC Curve C

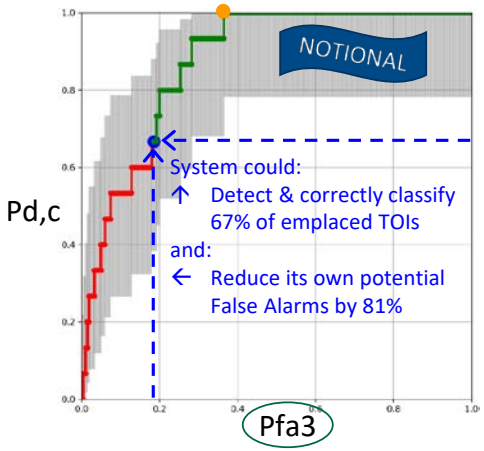
- Y Axis: $\text{Pd, c} = \frac{\text{\# True Alarms}}{\text{\# True TOIs}} \approx \frac{\text{\# True Alarms}}{\text{\# Emplaced TOIs}}$
- X Axis: $\text{FAR} = \frac{\text{\# False Alarms}}{\text{Test Area}}$
- **FAR** is a final count of the system's False Alarms after both Detection and Classification, normalized by the test area:
  - Pro: Easy to compare between demonstrations, since the False Alarm count is normalized by test area
  - Pro: Easy to compare between systems, since FAR is an absolute count of False Alarms



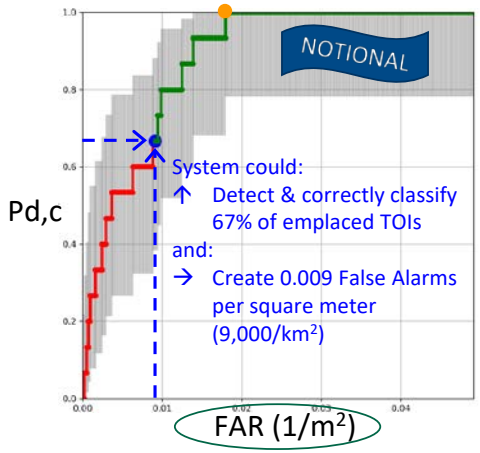
The solution is to construct a metric that accounts for the *area* of the test site. This metric, called False Alarm Rate (FAR), having units of inverse area, is just the number of False Alarms (the same numerator as before), now divided by the *Test Area*. This metric is easy to compare between demonstrations since test area is accounted for, and easy to compare between systems, since FAR is an absolute count of False Alarms similar to the FP metric.

A Tale of Two ROC Curves

ROC Curve A



ROC Curve C



Same system, same demo, same scoring → Same ROC curve shape
Different ROC curve axes → Different story to tell on system performance

The major point we'd like to leave you with is that there are many different scores that can be calculated from a demonstration, and each of those scores describes the system's performance from a slightly different perspective. The last several slides showed the same notional system from the same demonstration, scored with the same process, resulting in many ROC curves of the same shape—two of them are shown here. However, for each of those ROC curves, we plotted a slightly different metric on the X axis, which allowed us to interpret the system's performance from different perspectives. That is, we were able to tell different stories about how well the system could find TOIs and avoid false alarms:

ROC Curve A on the left, with Pfa3 on the X-axis, described how well the system's Classification step cleaned up the potential False Alarms it made during the Detection step—a *relative* measure of performance, comparing the system to itself. ROC Curve A is helpful when we're focused *solely* on the system's Classification step, compared to a *standard* detection method.

ROC Curve C on the right, with FAR on the X-axis, described how often the system, as a whole, created false alarms—an *absolute* measure of performance. ROC Curve C is helpful when we're focused on the system's Detection *and* Classification steps, *together*.

Both stories were legitimate, and both were fair assessments of the system. The question is: Which story do we want to tell for the UXO remediation community?

Questions

Jacob Bartel: jbartel@ida.org; 703 845 2172

Shelley Cazares: scazares@ida.org; 703 845 6792

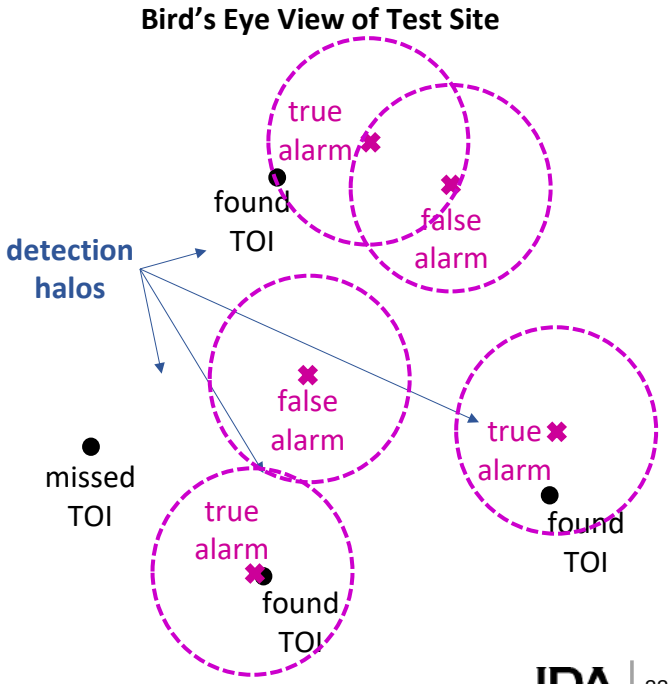
Please contact Jacob Bartel (jbartel@ida.org) or Shelley Cazares (scazares@ida.org) with any comments or questions.

Backups

The following slides may be useful in follow-on discussions.

Probability of Detection

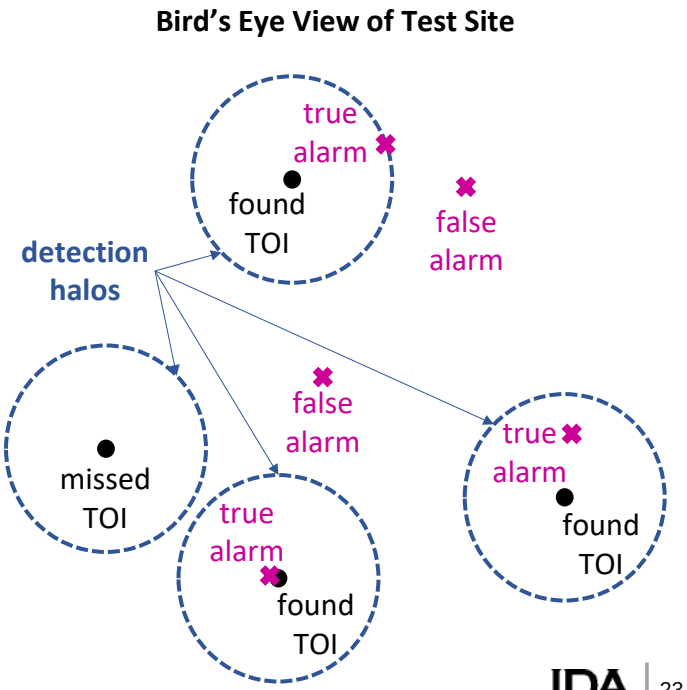
- Two types of missed TOIs:
  1. **TOI Miss Detection Error:**
     system fails to detect object when a TOI is actually there (example in figure)
  2. **TOI Miss Classification Error:**
     system detects TOI but mis-classifies it as Non-TOI
- **Both types** should be counted as False Negatives and included in the Pd,c metric



Since these tests are combined detection *and* classification tests, there are actually two ways to miss a TOI. First, a TOI Miss detection error, where the system fails to detect an object entirely. Secondly, a TOI miss classification error, where the system detects a TOI but mis-classifies it as a Non-TOI. Both of these types of errors are counted as False Negatives and included in the Pd,c metric.

False Alarm Rate

- Two types of false alarms:
  1. **False Alarm Detection Error:**
     system detects object even when no object is actually there (two examples in figure)
  2. **False Alarm Classification Error:**
     system mis-classifies Non-TOI as TOI
- **Only one type** (False Alarm Classification Error) should be counted as a False Positive and included in the Pfa metric

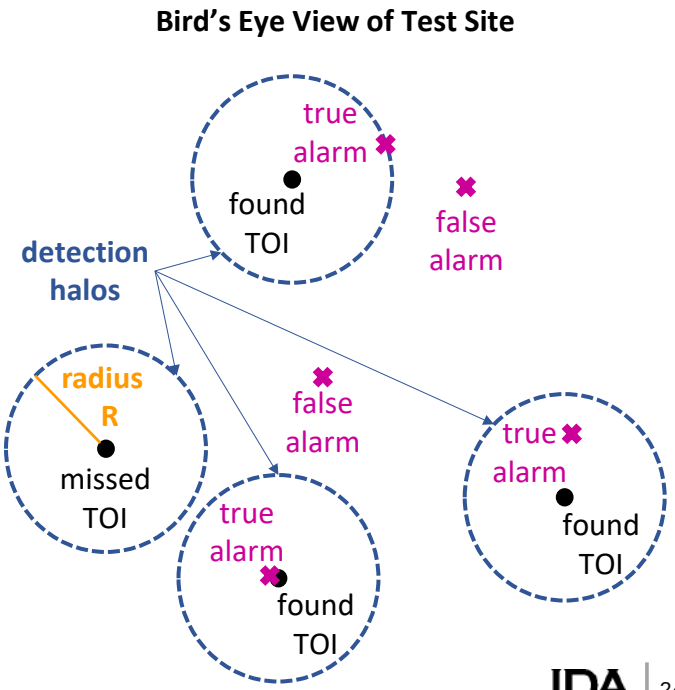


There are also two ways to make a False Alarm: A False Alarm detection error, where a system detects objects even when no object is actually there, and a False Alarm classification error, where a system mis-classifies a non-TOI as a TOI. However, only the False Alarm classification error should be counted as a False Positive and included in the FAR metric, to avoid double-counting False Alarm detection errors that are then passed on to the classification step.

Detection Halo Radius

- The detection halo is a circle centered around each true TOI
- Its radius R depends on:
  - Requirements of subsequent steps of remediation process (e.g., retrieval)
  - Geolocation error
  - Sensor resolution

Proper selection of R is *the* trickiest part of underwater demonstration scoring



We have frequently mentioned the ‘location’ of a TOI in the ground truth or the ‘location’ of a detection on a demonstrator’s call list. However, defining what this ‘location’ is has been one of the trickiest parts of scoring. The way that we define a ‘location’ of a TOI is by drawing an imaginary ‘halo’ around its ground truth coordinates with a certain radius (R). This radius R is based on various real-world parameters effecting the geolocation of TOIs—including: the requirements of the remediation process (or retrieval of the TOI), the geolocation error of both the ground truth and the UXO system, and the sensor resolution of the UXO system.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
March 2021	FINAL	

4. TITLE AND SUBTITLE

Metrics for Assessing Underwater Demonstrations for Detection and Classification of Unexploded Ordnance (Presentation)

5a. CONTRACT NUMBER

HQ0034-19-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Bartel, Jacob B.
Cazares, Shelley M.

5d. PROJECT NUMBER

AM-2-1528

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
Systems and Analyses Center
4850 Mark Center Drive
Alexandra, VA 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER

IDA Document NS D-21603

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

SERDP/ESTCP
Munitions Response
4800 Mark Center Drive, Suite 16F16
Alexandria, VA 22350-3605

10. SPONSOR/MONITOR'S ACRONYM(S)

SERDP/ESTCP

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited (19 March 2021).

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Receiver operating characteristic (ROC) curves are often used to assess the performance of detection and classification systems. The Strategic Environmental Research and Development Program / Environmental Security Technology Certification Program (SERDP/ESTCP) is sponsoring the development of novel systems for detecting and classifying unexploded ordnance (UXO) in underwater environments. SERDP is also sponsoring underwater testbeds to demonstrate the performance of these novel systems in relevant environments. The Institute for Defense Analyses is currently designing and implementing the scoring process for these underwater demonstrations to address the subtleties of ROC curve interpretation. This presentation provides an overview of the main considerations for ROC curve parameter selection when scoring underwater demonstrations for detecting and classifying UXO.

15. SUBJECT TERMS

acoustic detection; advanced geophysical classification (AGC); electromagnetic induction detection; environmental remediation; Receiver Operating Characteristic (ROC) curve; scoring; test and evaluation (T&E); underwater; unexploded ordnance (UXO)

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Bradley, David
Unclassified	Unclassified	Unclassified	Same as Report	55	19b. TELEPHONE NUMBER *(Include area code)* (571) 372-6388

Reset